Error correction on a tree: An instanton approach

V. Chernyak [a], M. Chertkov [b], M. G. Stepanov [b,c,d], B. Vasic [e]
[a] *Corning Inc., SP-DV-02-8, Corning, NY 14831, USA*
[b] *Theoretical Division, LANL, Los Alamos, NM 87545, USA*
[c] *Department of Mathematics, University of Arizona, Tucson, AZ 85721, USA*
[d] *Institute of Automation and Electrometry, Novosibirsk 630090, Russia*
[e] *Department of Electrical Engineering, University of Arizona, Tucson, AZ 85721, USA*
(Dated: March 25, 2004)

We identify a family of phase transitions found for post-error-correction Bit-Error-Rate (BER) of finite-size Low Density Parity Check (LDPC) code approximated by a tree-like structure. The problem of testing an LDPC code performance is re-formulated in terms of statistical mechanics as an intergal over noise realizations on a graph. The integral is approximated by the sum over different symmetry saddle-point solutions (instantons). Different phases, related to different values of the Signal-to-Noise-Ratio, correspond to different instanons that dominate the integral for BER.

PACS numbers: 89.70.+c,05.20.-y Submitted to Phys.Rev.Lett. 03/24/04

Last fifty years have witnessed a tremendous increase of amount of data being transmitted through various communications systems. Although data flow through these systems has increased tremendously, a major concern is the ability to ensure error-free data transfer in the presence of noise and other impairments during transmission. The problem of dealing with errors in information transmission has fundamental importance and has been studied extensively in information theory and coding theory. In 1948 Shannon [1] has proved that applying an Error-Correcting Code can result in an error-free communication in the thermodynamic limit of an infinitely long word, as long as the rate of transmitted information is kept below the channel capacity. Constructing good and practical capacity-approaching codes has been a challenge until the discovery that long Gallager codes [2] can achieve near-optimum performance when used for transmission over white additive Gaussian noise channels [3]. In the past few years several codes were designed with performances very close to this limit [11]. Generally, these codes are referred to as codes on graphs, and their prime examples are Low Density Parity Check (LDPC) codes and turbo-codes. Also, a significant insight into iterative decoding was gained due to its interpretation in terms of message passing and belief propagation in graphical models. Most of LDPC codes are based on random constructions of the coding maps, however, it has been also shown [4] that regularly structured LDPC codes (that have a natural advantage of being memory effective and simpler to build) can be performing comparably well. Another major recent development is associated with reformulating the error-correction problem in terms of statistical mechanics [5], which stimulated a fresh flow of analogies and new exciting ideas and analogies. (See e.g. [6–8].) However, the new approach has mainly focused on comprehensive analysis of the thermodynamic (infinite code length) limit, whereas describing the phenomena related to realistic finite-size codes has attracted much less of attention.

Performance of any finite-size error-correcting code is measured in terms of the dependence of the post error-correction Bit-Error-Rate (BER) on the Signal-to-Noise Ratio (SNR). Error correction aims to decrease the BER by adding redundant information (overhead) to information messages. The smaller the post error-correction BER is for fixed overhead, the better. Any new generation of communication devices creates a new challenge for the error-correction technology as it sets higher standards for the channel capacity thus lowering the level of BER which can still be tolerated. Straightforward Monte Carlo numerical simulations constitute an efficient method only for the values of BER $\sim 10^{-7}$ or higher, and it falls short in accessing lower values of BER. Experimental tests are extremely expensive, thus frequently impractical, since they require building a special device prototype for any new suggested coding/decoding strategy. This implies that finding efficient practical ways of extremely low BER evaluation is under universal demand. Our main objective is constructing a theoretical tool capable of delivering quantitative estimates for these low probability events analytically. The approach we propose to adopt and develop for achieving this goal is known under the names of saddle-point, optimal fluctuation, or instanton calculus. This method, aiming to estimate a low probability event, is common in modern theoretical physics, introduced initially in the context of disordered systems [9].

The letter is organized as follows. We start with a general and brief introduction to the subject: We describe the basic principles of coding for an LDPC code, introduce the optimal Maximal-A-Posteriori (MAP) decoding strategy along with generally suboptimal yet very efficient Belief Propagation (BP) decoding, and finally define the post-error correction BER that characterizes the code performance. Next we argue, following [3, 7, 8, 10, 11], that a finite-size tree-like structure offers

a good approximation for an LDPC code if the length of the shortest loop on the corresponding Tanner graph is long enough. We further focus on the BER computation for the central site on the tree, presenting it as an integral over noise configuration (fields) on the tree. Instantons – special configuration of the field giving the major contribution into the integral – are first found numerically through complete variational procedure. It is shown that all the relevant instantons, all of different symmetries, can be characterized in terms of partially colored Tanner graph. Finally we describe the main result of this letter, namely a sequence of phase transitions, between phases/instanotns of different symmetries, emerging for BER with the SNR change.

In the case of binary linear block coding the error correction consists of: (1) coding the original message (word) represented as a set of L Boolean, ± 1, symbols into a longer word consisting of N Boolen signals; (2) transmitting the N bit long codeword through a noisy channel; (3) decoding the corrupted message detected at the output. Any binary linear code can be conveniently described by its Tanner graph, consisting of N variable nodes (marked by Latin indices) that correspond to the bits of the transmitted message and $M = N - L > 0$ checking nodes (marked by Greek indices) that represent the parity checks, and the connections occur between those bits j and parity checks α so that the bit j participates in the parity check α, i.e. $j \in \alpha$. (In this representation all the parity checks should be linearly independent.) More formally, $\boldsymbol{\sigma} = (\sigma_1, \cdots, \sigma_N)$ with $\sigma_i = \pm 1$ represents one of 2^L code words if and only if $\prod_{j\in\alpha} \sigma_j = 1$ for all the checking nodes, $\alpha = 1, \cdots, M$. The code redundancy is described by the overhead $M/L = R^{-1} - 1$, with $R = L/N < 1$ being the code rate. Transmitted through a noisy channel a code word gets corrupted due to the channel noise, so that at the channel output one detects, $\boldsymbol{x} \neq \boldsymbol{\sigma}$, where in the simplest model case of the additive white Gaussian channel considered here $\boldsymbol{x} = \boldsymbol{\sigma} + \boldsymbol{\varphi}$, $\langle \boldsymbol{\varphi} \rangle = 0$, $\langle \varphi_i \varphi_j \rangle = \delta_{ij}/s^2$, where s measures the SNR.

The goal of decoding is restoring the best approximation for the original message from a corrupted word. Optimal decoding, also known under the name of Maximal A Posteriori (MAP) Symbol decoding, can be represented in terms of the generating function of an effective "spin" model: $\exp[-F(h)] = \sum_{\boldsymbol{\sigma}} \prod_{\alpha=1}^{M} \delta\left(\prod_{j\in\alpha} \sigma_j, 1\right) \exp\left(\sum_{k=1}^{N} h_k \sigma_k\right)$, where the "external magnetic field" $\boldsymbol{h}$ is related to the channel noise $\boldsymbol{\varphi}$, $\boldsymbol{h} = s^2(1 + \boldsymbol{\varphi})$, $\delta(x, 1)$ is the Kronecker δ-symbol and the "magnetization", defined as $\psi_j(\boldsymbol{h}) \equiv \langle \sigma_j \rangle = -\partial F(h)/\partial h_j$, is interpreted as the result of decoding, or more accurately sign$[\psi_j]$ gives the decoded value for the bit j. The code performance can be characterized via the density of errors at the given site j known as the post-error-correction BER that can be also described as the probability of a spin flip:

$$B_j = \int_{-1}^{0} d\zeta \int d\boldsymbol{h} \;\; \delta\left(\psi_j\{\boldsymbol{h}\} - \zeta\right) \prod_{j=1}^{N} f(h_j), \qquad (1)$$

where $f(x) \equiv \exp[-(x - s^2)^2/(2s^2)]/\sqrt{2\pi s^2}$ and $\boldsymbol{\sigma} = \mathbf{1}$ is assumed for the codeword input. (Due to the linearity and homogeneity of the code with respect to the permutation of the code words, BER defined for any other initial codeword would be exactly equal to the one given by Eq.(1).)

MAP decoding is optimal, however, inefficient, since it requires an exponentially large number of 2^L steps. BP decoding [3, 10] constitutes a fast (linear in N) yet generally approximate alternative, corresponding to replacing generating function in MAP by solving the following set of nonlinear equations (hereafter referred to as the BP equations) $\eta_{j\alpha} = h_j + \sum_{\beta\neq\alpha}^{j\in\beta} \tanh^{-1}\left(\prod_{i\neq j}^{i\in\beta} \tanh(\eta_{i\beta})\right)$, $\eta_j = h_j + \sum_{\beta}^{j\in\beta} \tanh^{-1}\left[\sum_{i\neq j}^{i\in\beta} \tanh(\eta_{i\beta})\right]$, where $\tanh^{-1}(\psi_j) \equiv \eta_j$. Iterative solution of the BP equations truncated at a finite step is known as Message Passing (MP) algorithm. As shown in [10] the set of BP equations becomes exactly equivalent to MAP in the loop-free approximation. Using physics jargon, it is equivalent to the Bethe-lattice approximation [12]. This basic approximation involves generating a tree with the number of generations, counted from the central variable node to be equal to the shortest loop length on a realistic graph. Note that for Gallager codes the typical length of the shortest loop is estimated as $\sim \ln N$ [11]. Although the method of BER computation proposed in this letter is generally applicable for any kind of codes, we will focus solely on the regular codes for which each variable node participates in $m \geq 2$ checking

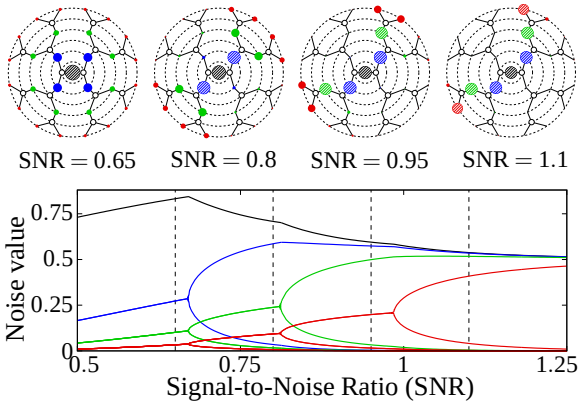

FIG. 1: $m = 2, l = 3, n = 3$. Symmetry instantons and bifurcation picture for complete optimization procedure (no symmetry was a-priori) assumed. On the first line area of a circle surrounding any variable node is proportional to the value of the noise on the node. Different colors correspond to different generations on the tree.

node, and each checking node constraint includes $l \geq 3$ variable nodes, with $l > m$.

The set of the δ-functional BP constrains, leads to essential complication in the generic case resulting in a nontrivial statistical mechanical model. However, in the tree-like case (no loops) the constraints become fairly easy to handle. Indeed, in this case each variable site can be described by one "inbound" field $\eta_{j\alpha}$ with the checking site α belonging the only path from the given variable site to the tree center, and the other $m-1$ "outbound" fields $\eta_{j\beta}$ with $\beta \in j$ and $\beta \neq \alpha$. It is a remarkable feature of the tree structure that the integrand in Eq.(1) can be expressed solely in terms of the "inbound" fields on the tree and only "outbound" field defined exactly in the center of the tree. Therefore, the only nontrivial integrations go over the "inbound" fields, hereafter denoted by simply η_j, and Eq.(1) is simplified: $B_0 = \int_{-1}^{0} d\zeta P_0(\zeta) \sim P_0(0)$ and $P_0(\zeta) = \int (\prod_j d\eta_j) \exp(-\mathcal{Q})$, where the effective action is

$$\mathcal{Q} \equiv \frac{1}{2s^2}\left(\zeta - \sum_{\beta\in 0} \tanh^{-1}\left(\prod_{k\neq 0}^{k\in\beta} \eta_k\right) - s\right)^2 + \frac{1}{2s^2}\sum_{j\neq 0}\left(\eta_j - \sum_{\beta\in j} \tanh^{-1}\left(\prod_{k\neq j}^{k\in\beta} \eta_k\right) - s^2\right)^2, \quad (2)$$

and $j = 0$ marks the tree center, $\beta > j$ denotes that the check node β is positioned above the variable node j in the tree hierarchy.

Integrations over noise fields η_j will be performed in the saddle-point instanton fashion, that corresponds to the assumption that the major contribution to the integral originates from the special (instanton) configurations related to the minimum of the effective action $\mathcal{Q}$: $\delta\mathcal{Q}/\delta\eta = 0$. Alternatively, one can solve the BP equations on the tree using the MP algorithm (i.e. making some fixed number of iterations), substitute it into the resulting expression for the magnetization/BER, and maximize it with respect to the noise field. The two variational schemes should be equivalent in the limit of infinite number of iterations in the MP case (we found the convergence with the number of iterations to be relatively fast and monotonic in the loop-free case). The result of the MP variational procedure for $m = 2$, $l = 3$, the number of generations on the tree $n = 3$, and 10 iterations is shown on Fig. 1. Full variation over all noise fields on the tree (thus containing no symmetry assumption) shows rich bifurcation picture corresponding symmetry breakdown. At small values of SNR the optimal solution is of maximal symmetry with all noise fields that belong to a given generation (counted from the tree center) being identical. With SNR increase the symmetry of the optimal configuration degrades discreetly through n steps. The symmetry of the k-th order instanton can be described by a set of variable nodes (marked striped on Fig. 1) that extend from the center (which is always striped/marked) towards the k-th generation according to the following rule: All checking nodes connected to a marked variable node of previous generation are marked, while for any marked checking node exactly one variable node of the next generation is marked. The rule is generic, i.e. it applies for any values of m and l.

Taking the symmetry assumption as granted one can substantially simplify and improve the process of finding the set of instanton solutions and getting a better estimate for BER. Thus the independent fields that correspond to an instanton with the symmetry broken up to the k-th order can be conveniently represented in terms of the two-index quantities $\eta_j^{(p)}$ using the following agreement. The variable $\eta_j^{(p)}$, where $p = 0, \cdots, k$ and $j = 0, \cdots n-1-p$, represents the field on a non-marked node located in generation j (counting from the leaves), so that the first marked node on the only path to the center lies in generation p (counted from the tree center). The variable $\eta_{n-p}^{(p)}$ with $p = 1, \cdots, k$ represent the field on a marked node that is located in the generation p (counting from the center). Replacing the full set of the η-fields on the graph by the described above restricted symmetry set $\{\eta_j^{(p)}\}$, substituting it into the effective action $\mathcal{Q}$ described by Eq.(2), and minimazing the resulted k-th order effective equation with respect to the k-th order restricted set of η fields one arrives at a system of equations that the k-th order instanton that are bulky and are not presented here. The set of equations for the k-th instanton can be formulated in terms of a $k+1$-dimensional minimization problem. We have found, however, that the system can be approximately reduced to a one-dimensional chain minimization problem if either of the following conditions holds: (i) $l \gg 1$; (ii) $n, n-p \gg 1$ and $s > s_c$, where s_c is defined as such s which formally solves the system, $\eta = g(\eta)$ and $1 = g'(\eta)$ where $g(\eta) = s^2 + (m-1)\tanh^{-1}\left(\tanh\left(\eta^{l-1}\right)\right)$; (iii) $s \gg s_c$. Note, that in the thermodynamic limit action of the high-symmetry instanton, which is finite at $s < s_c$ becomes infinite at $s > s_c$, with s_c being finite for $m > 2$. In all three cases the instantons have the following structure: The unmarked variables $\eta_i^{(p)}$ with $p > 0$ grow while approaching the center according to the equation $\eta_j^{(p)} = g(\eta_{j-1}^{(p)})$, whereas for the marked variables $\eta_{n-p}^{(p)} \approx 0$. Therefore, the only dynamical field to be optimized is the unmarked portion of the $p = 0$ branch. Note, that although the approximation is justified only in either of the three aforementioned limits, it actually works quantitatively well even for the moderate values of the key parameters l, m, n, s, as follows from comparing the numerical solutions of the full (i.e. making no a-priori symmetry assumptions), $k+1$-dimensional and approximate 1-dimensional minimization problems.

Within the instanton approximation BER is estimated as $B_0 \sim \sum_{k=0}^{n} N_k \exp[-\mathcal{Q}^{(k)}] c_k$, with $\mathcal{Q}^{(k)}$ being the ac-

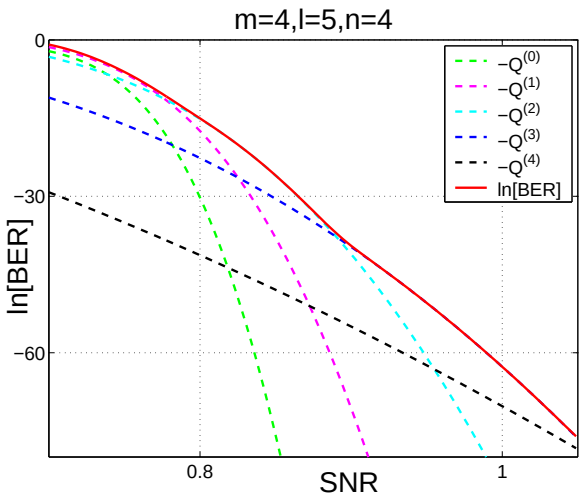

FIG. 2: $m = 4, l = 5, n = 4$. Comparative plots of BER (full sum, but the phase volume factors were not counted: $c_p \approx 1$) and individual instanton contributions, calculated within the single-chain approximation, vs SNR.

tion of the k-th order instanton. The combinatorial factor N_k with $N_0 = 1$ and $N_k = m(m-1)^{k-1}(l-1)^k$ accounts for the symmetry-induced instantons degeneracy. The phase space volume c_k occupied by a given instanton accounts for Gaussian fluctuations in the neighborhood of the instanton solution. Calculating c_k that constitutes an important yet difficult task is postponed for future studies. In any of the three asymptotic limits (i-iii) mentioned above the sum in the definition of BER is dominated by a single instanton contribution that determines the relevant SNR phase. Moreover, in either of the three limits the value of the instanton action dominates the contribution, with both the combinatorial N_p and the phase space c_p factors related contributions being sub-leading. For the lowest SNR values the major contribution to BER originates from the most symmetric instanton. With the SNR increasing the system experiences a series of phase transitions from $\mathcal{Q}^{(0)}$ to $\mathcal{Q}^{(1)}$, $\mathcal{Q}^{(2)}$, etc. to $\mathcal{Q}^{(n)}$ that take place at $s_1 < s_2 < \cdots < s_{n+1}$ respectively. Note, that at $n \to \infty$ an infinite sequence of s_k, with $k < n$, converges to s_c from below. In the case of a finite tree shown in Fig. 2 the transitions are not that sharp (especially those corresponding to relatively low values of SNR), yet still recognizable.

Emergence of the sequence of instantons/phases/transitions reported above can also be understood intuitively: If the noise is large correlations between the noise values on different nodes are weak, thus no symmetry breaking (marked) structure on the Tanner graph is possible and therefore the most symmetric noise configuration is optimal. The correlation length growth due to the SNR increase leads to developing a preferred/marked structure that breaks the full symmetry. The structure grows from the tree center towards the leaves, simply because the tree center is chosen for the local measurement of BER. In the extreme case of large SNR the symmetry brake down is obviously associated with the structure of the code word closest to the original one, thus making the logarithm of BER to be proportional to the Hamming distance between the two special codewords, and also rationalizing why (for any instanton solution) the marked structure locally resembles the structure of the next to original code word. Note also, that emergence of a finite correlation length (on the graph) growing with the SNR increase suggests that the tree approximation works well for a finite LDPC code as long as the correlation length is short compared to the length of the shortest loop on the LDPC graph. Thus, the no-loops/tree approximation is perfectly justified for at least some number of low SNR phases. For the higher SNR phases the approximation may still be reasonable, however, resolving this challenging question requires going beyond the tree approximation. We conclude with noting that emergence of the sequence of transitions suggests a substantial flattening of the BER dependence on SNR at moderate values of the latter. This observation may have an interesting relation to the error floor phenomenon reported for the Frame (code word) Error Rate [13], and that the "near codewords", which are claimed to be giving the major contribution to the error floor phenomenon [14], are reminiscent of the instantons with partially broken symmetry.

We are thankful to I. Gabitov for many fruitful discussions and support, we also acknowledge very useful comments of D. Sherrington and A. Montanari that stimulated the development of the project on its early stages.

[1] C.E. Shannon, Bell. Syst. Tech. J. **27**, 379 (1948).
[2] R.G. Gallager, *Low density parity check codes* (MIT Press, 1963).
[3] D.J.C. MacKay, IEEE Inf. Theory **45**, 399 (1999).
[4] B. Vasic, I.B. Djprdjevic, R.K. Kostuk, JLT **21**, 438 (2003).
[5] N. Sourlas, Nature (London) **339**, 693 (1989).
[6] A. Montanari, Eur. Phys. J. B **23**, 121 (2001).
[7] J.S. Yedidia, W.T. Freeman, Y. Weiss, www.merl.com/papers/TR2001-16/ .
[8] R. Vicente, D. Saad, Y. Kabashima, Eur. Lett. **51**, 698 (2000).
[9] I.M. Lifshitz, Usp. Fiz. Nauk **83**, 617 (1964).
[10] J. Pearl, *Probabilistic reasoning in intelligent systems: network of plausible inference* (Kaufmann, 1988).
[11] T.J. Richardson, R.L. Urbanke, IEEE Inf. Theory **47** (2) 599 (2001).
[12] H.A. Bethe, Proc.Roy.Soc.London A, **150**, 552 (1935).
[13] C. Di, D. Proietti, I.E. Telatar, T.J. Richardson, R.L. Urbanke, IEEE Inf. Theory **48**, 1570 (2002).
[14] T. Richardson, "Error floors of LDPC Codes", 2003 Allerton conference Proccedings.